CAMBRIDGE ASSESSMENT

Equating methods used in KS3 Science and English

Paper for the NAA technical seminar, Oxford, 23-24 March 2006

Tom Bramley
Research Division
8th March 2006

The opinions expressed in this paper are those of the author and should not be taken as official policy of Cambridge Assessment or any of its subsidiaries.

UNIVERSITY *of* CAMBRIDGE
Local Examinations Syndicate

Contents

Introduction

There are four separate sources of evidence which are taken into consideration when determining the location of the cut-scores on NC tests:

1. Statistical equating using pre-test data.
2. Judgmental exercises using practising teachers.
3. Scrutiny of 'live' scripts by panels of senior markers. This exercise takes place after the test has been marked, but before the cut-scores have been finally confirmed.
4. 'Impact data' using the live test raw score distribution from a sample of around 20 000 pupils - this gives a very good indication of the consequences of each set of proposed cut-scores on the proportion of pupils who will achieve each level.

These four methods are not independent – the statistical cut-scores derived from (1) might play a part in some of the judgmental methods (2); and the results of (1), (2) and (4) determine the mark ranges of live scripts considered in (3). It is methods used in the first of these (statistical equating using pre-test data) which are the subject of this paper.

The authors of the papers for this seminar were encouraged to present a personal perspective on the issues and what follows is based on my experiences working in the area of National Curriculum testing – occasionally from 1996 to 2000, and full time from 2002 to 2005[1]. The first section describes some of the conceptual issues, the second section describes some of the historical background, the third section describes the current methods and the final section raises some issues for debate.

1. Conceptual issues in equating National Tests

The first point to note about statistical equating in the context of setting level cut-scores on National Tests is that it is a standard-maintaining exercise, rather than a standard-setting one. (See, for example, Pollitt (1994), Bramley (2005a)).

The core assumption underlying all the (statistical) methods we have used for standard-maintaining is that the task is essentially a psychometric problem. On this viewpoint:

"… a test score is seen as an indicator of a pupil's position on an abstract or 'latent' trait (Lord, 1952). This trait, sometimes referred to as a variable or construct or dimension, is conceived as a continuum analogous to physical dimensions like length or temperature. The standard is the point on this latent trait that marks the boundary between two meaningfully different categories (e.g. pass/fail, A/B, level 5 / level 4). This standard exists independently of the test itself and therefore has to be applied to each new test that is constructed. … on National Curriculum tests, the aim is to ensure that the standard set in previous years is applied to the current year's test when setting the cut-score at each level." (Bramley, 2005b)

It follows from this approach that a pre-requisite for any equating is that the latent trait is the same – in other words that the tests being equated are measuring 'the same thing'. We have to assume that it is meaningful to say that a level 5 on one year's test 'means the same' as a level 5 on another year's test. (We will see later that one problem which often arises is that of changes to the test specification or to the assessment model.) It is assumed that every child in each year cohort is at a particular position on the latent trait. As a shorthand, we refer to this position as their 'ability' in the subject, but this does not mean some innate ability or IQ. It merely refers in a neutral psychometric sense to their level on the trait being measured.

[1] The contract to develop the KS3 tests in English and Science has been held by UCLES since 1993. The test development team was part of the Research and Evaluation Division from 1993 until 2004, when it moved to OCR.

Because tests inevitably differ in difficulty, no matter how carefully designed to a specification (and NC tests are very carefully designed to a detailed test specification) a given raw score on two tests will not necessarily correspond to the same level of ability. An informal way of expressing the goal of equating in the NC context is that it is "trying to ensure that pupils with the same level of ability will be awarded the same NC level on the current year's test that they would have been awarded on a previous year's test." (Bramley, 2005b). This means that the emphasis is on equating the mark points which correspond to the level cut-scores, rather than every single mark on both tests. However, the statistical methods which are used do in fact usually provide an 'equating function' which can map each raw score on one test to an equivalent raw score on another test.

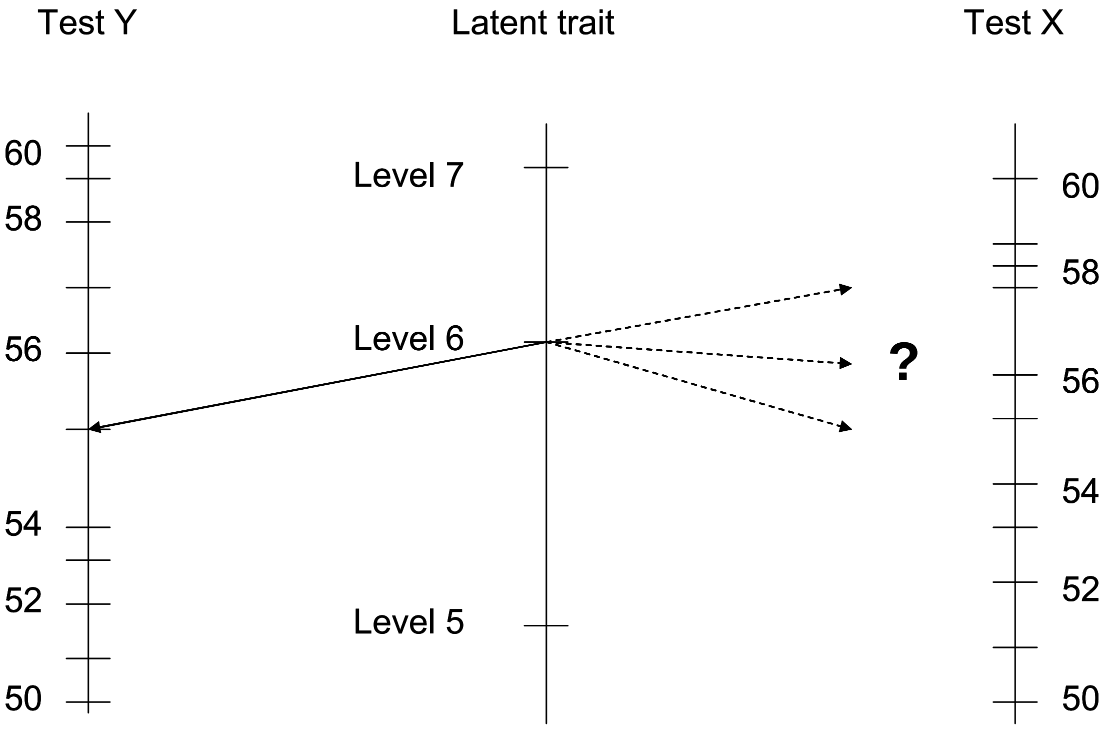

Figure 1: Illustration of equating for the level 6 cut-score[2].

Figure 1 shows a portion of the raw mark scales of two tests, X and Y, and the latent trait, on which three NC levels (5, 6, and 7) are marked. The arrow on the left denotes a mark of 55 corresponding to the Level 6 cut-score on test Y. The task of equating is to find the corresponding mark on test X.

The visual representation of Figure 1 invites an analogy of the latent trait with a physical dimension such as length. Before the days of the digital display, the output of many physical measurement instruments was the displacement of a needle by a distance corresponding to the magnitude of the quantity measured. Three examples are bathroom scales (weight), aneroid barometers (air pressure), and voltmeters (electrical potential difference).

However, this tempting analogy should be treated with some caution, for several reasons:

1. The 'ontological status' of physical variables (i.e. whether they are 'real' or not) such as length or weight is not as disputable[3] as the ontological status of latent psychological variables (Borsboom et al, 2003).

[2] This diagram resembles Figure 1 in Newton (2005). However, he was explaining the term 'linking construct', which has a different meaning from what I intend to convey by 'latent trait'.

2. Psychological variables are often thought to be multi-dimensional – for example a test of 'general academic ability' might contain questions requiring a number of different skills and knowledge of a number of different topic areas. Most (all?) physical variables are unidimensional.
3. Michell (2000) insists that whether a variable possesses a quantitative structure is an empirical question which should be answered before any attempt to measure that variable. On this hard-line view it is possible that psychological variables such as 'KS3 Science ability' do not have the necessary quantitative structure to allow them to be measured.
4. Measurement error in physical measurement instruments is often small compared to the magnitude of the quantity being measured. With latent psychological variables the measurement error is relatively larger, and (somewhat paradoxically) is an essential feature of some of the mathematical models used to estimate the measures (see, for example, Pelton & Bunderson, 2003).
5. Physical measuring instruments are usually designed on the basis of an explicit theory about both the nature of the quantity being measured, and how it interacts with the instrument to produce the reading. In the case of temperature, for example, this might be the theory of how mercury expands when heated. This kind of explicit theory is often absent, or only vaguely articulated, in the case of latent psychological variables (Stenner, 1996). With physical measurements, such theories usually provide a means of ensuring that equal displacements of the needle on the display correspond to equal amounts of the quantity being measured.[4]
6. Physical variables usually have an agreed (arbitrary) unit of measurement. For example, temperature can be measured in degrees Celsius or degrees Fahrenheit. See the next section for some discussion of the measurement unit in KS3 English.

These factors all caution against a too-hasty assumption that the raw scores on two tests can be treated in the same way as readings from two physical measurement instruments designed to measure the same quantity. Figure 1 attempts to encourage this attitude of caution by not having equal spaces either between the levels on the latent trait, or between the raw marks on the two tests. However, this only addresses points 5 and 6 above. The first four, more philosophical, points are still ignored in Figure 1.

This may seem like a laborious re-iteration of what everyone knows. But when the outcomes of statistical equating are unsatisfactory (for whatever reason) the test development agencies often hear complaints like 'why can't they get it right?' or 'why do they all use different methods?' or 'why do the different methods all give different results?' or 'why does it have to be so complicated?'. The implication is that it is something to do with the choice of method for equating. This may be partly true, but (in my view) the greater part of the problem is likely to be related to points 1 to 6 above.

[3] Although in the branches of philosophy which deal with these things, most things are disputable, as might be expected!

[4] Of course, it is sometimes the case that the instrument reports on a scale related to the quantity by some mathematical transformation such as the logarithmic transformation (used, for example, in reporting sound level in decibels).

2. Historical background

Because the KS3 Science and English tests cover very different subject domains, with different assessment models, and have been developed by different teams[5], it is easier to describe historical developments for each subject separately.

2.1 English

Equating rationale
The theoretical basis for maintaining standards on the KS3 English test by statistical equating was described in detail by Pollitt (1994). He noted that the unit of the 'manifest scale' (the reported level) was originally defined in the TGAT[6] report to be equal to two years of progress, but that performance and content descriptors for each level were then created without recourse to empirical data about what children at different ages could do – effectively replacing the TGAT definition with a new manifest scale "… with units that are no longer obviously equal interval with respect to time, or ability, or anything" (Pollitt, op cit).

However, he argued that since the public would assume that the level bands would in some sense be equal in size, they should be made equal in terms of *equal amounts of English abilit*y. In other words, on the latent trait in Figure 1, the levels should be equally spaced. This then led naturally into the choice of the Rasch model as the appropriate statistical tool for equating, since it yields estimates of person ability and item difficulty on an equal-interval logit scale (see, for example, Wright ,1977). The logit scale can be re-scaled by any linear transformation to produce a more readily interpretable scale. This involves fixing a point on the scale, and a unit size. They used empirical data from a large sample (900) of KS3 pupils to set the scale in 1993, by setting the average ability of their sample to a scale score of 60 (which should have corresponded to the level 5/6 boundary on the TGAT definition), and the unit at a tenth of a level (based on information about the national distribution of levels in 1993 and Teacher Assessed levels of the pupils in their sample). Thus scale scores of 50-59 corresponded to the level 5 band, and scale scores of 60-69 to the level 6 band etc.

Equating procedure
Of course, the fact that the KS3 English test contains questions which require more subjective judgment on the part of the marker than do the KS3 maths or science tests creates an extra element of uncertainty in the results from any equating method. But the procedure for equating using this approach was simply a case of calibrating each new test onto the Rasch scale as defined above. This ensured that differences in difficulty between different tests could be taken into account – the 'conversion table' for each test would show the raw scores corresponding to the scale scores. For an easy test, a higher raw score would be needed to obtain a scale score of (say) 60 than on an easier test.

However, though simple in principle, in practice there were several difficulties with this approach:

- Anchoring – in order to link together the scales created from tests containing different questions taken by different pupils it was necessary to use an anchor test. The earliest incarnation of this was a comprehension passage with some short-answer questions, followed by a passage of text with a cloze (gap-fill) summary. This was revised in the late 1990s when the comprehension passage with short-answer questions was replaced with some multiple-choice verbal reasoning items (making the anchor test consist solely of objective items, which could be more quickly and more reliably marked). Both of these anchors were different in nature to the KS3 test questions, and tended to 'misfit' severely when analysed with the Rasch model.

[5] Obviously the questions and other test materials were developed by different teams, but before 2002 within UCLES the statistical evaluations were also carried out by different staff.
[6] Task Group on Assessment and Testing. Department of Education and Science, 1987.

- Lack of flexibility - the method ensured that each new test was linked by the same anchor to the original 1993 scale. This was in fact good practice, because it prevented the ‘drift’ in equating which can occur if each year’s test is equated to the previous year’s. However, in the politically pressured environment of level setting, it began to produce cut-scores which were unacceptable. The fact that the method constrained each level to be equal interval with respect to ability meant that if final live cut-scores were set which did not meet this requirement, the method was guaranteed to produce new cut-scores which did not match the latest actual live ones. Re-calibrating the scale to get one level the ‘correct’ width would automatically have meant that the other levels would be ‘wrong’.
- Unpopularity of Rasch – there had been a long-running debate / controversy in the UK about the validity of using Rasch (or even other IRT methods) for trying to maintain standards over time. The relative statistical complexity of the method, and the relative lack of technical expertise among many of the staff at the client agency (SCAA / QCA), and the fact that other agencies were using different methods for test equating meant that the decision-makers at SCAA / QCA were in the unenviable position of having to ‘take on trust’ much of our analysis, and we were in the unenviable position of having to defend it with many tortuous caveats and complex explanations. Most of the research literature on Rasch and IRT concentrates on tests containing dichotomous (1-mark) items, or rating scales (e.g. questionnaires with 5 response categories). There was therefore little or no available literature to refer to when presenting our analyses.
- The KS3 English test itself – it would be an understatement to say that the specification for this test was not especially conducive to Rasch or any other statistical method[7] of equating! From 1995 to 2002 the test consisted of two papers. Paper 1 contained three reading comprehension questions (worth 11, 6 and 11 marks) based on two texts with a linked theme followed by a choice of one of three writing tasks, also on the same theme (worth 33 marks, but only even-numbered marks from 0 to 30 could be allocated, then 31, 32 or 33!). Paper 2 was the Shakespeare paper. Schools would prepare their pupils on one of three Shakespeare plays. Paper 2 contained a choice of one from two tasks per play. The response to this task was marked twice, once for ‘Understanding and Response’ (22 marks) and once for ‘Written Expression’ (16 marks). Such a structure seemed unlikely to meet the requirement for Rasch (or other IRT) models of unidimensionality and local independence (of item responses).

We attempted to address the first of these problems in the period 2000-2002 by using the anchor as a separate test whose purpose was merely to estimate the mean and SD of English ability in the pre-test sample, rather than as a true anchor forming part of the Rasch analysis. However, this was not sufficient to counter the other problems and in 2003, when the test model changed, we changed our equating method (see section 3).

2.2 Science

As noted earlier, the KS3 science test was developed by a different team within UCLES with different statistical staff. Rasch was not used at all (somewhat ironically, given that the KS3 science test, with a large proportion of dichotomous items, would seem more likely to meet the requirements of the Rasch model).

The KS3 science test has two tiers – the lower tier (T3-6, 180 marks) contains questions targeting levels 3 to 6, and the higher tier (T5-7, 150 marks) targets levels 5 to 7[8]. The questions at levels 5 and 6 (90 marks’ worth) are common to both tiers. Each mark on the test can be attributed to a particular level, and to a particular content area (one of the four ‘attainment

[7] The fact that the mark scheme was level-related meant that in theory it was possible to derive cut-scores, or at least possible ranges for the cut-scores, purely from the mark scheme.

[8] The ‘extension test’ targeting levels 8 and ‘EP’ (Exceptional Performance) was phased out in 2002.

targets' Sc1 – Sc4[9]). In the early days of the test a 'criterion-related' method was proposed for setting the cut-scores (Massey, 1995). Very briefly, this would have involved specifying a level cut-score in terms of expectations about percentage of success on the questions at each level. For example, Massey suggested the following scheme for setting the cut-scores on Tier 3-6 paper:

Level 3: 60% of marks available at L3 + 30% at L4 + 20% at L5 + 10% at L6
Level 4: 70% at L3 + 60% at L4 + 30% at L5 + 20% at L6
Level 5: 75% at L3 + 70% at L4 + 60% at L5 + 30% at L6
Level 6: 80% at L3 + 75% at L4 + 70% at L5 + 60% at L6.

A scheme like this could (in principle) be implemented without any pre-testing or statistical equating, although Massey strongly advocated pre-testing and statistical evaluation of items to check that they were indeed targeting the intended levels. Note that this scheme does not specify how pupils should actually gain their marks in the live test in order to be awarded a level, but specifies an algorithm for calculating the cut-scores based on the mark profile of each test.

As far as I am aware, this approach was neither used to set the cut-scores when the tests became externally marked in 1995, nor been used since. Although it is extremely transparent and easy to explain to teachers and the public, I suspect that this method was not adopted because of the danger of it producing a politically unacceptable outcome in terms of percentage pass rates. The only ways to modify the cut-scores generated from this procedure are by changing the (arbitrary) percentages in the algorithm, or by re-classifying questions to change their level, both of which would be difficult to justify on any *a priori* grounds.

So, in practice, the statistical equating for KS3 science was carried out using one or more of several standard equating designs and methods, as described in Petersen et al. (1989). All of these methods involve matching the distributions of two sets of test total scores in order to obtain an equating function linking raw scores on one test to equivalent raw scores on the other test. Linear equating methods involve matching just the mean and standard deviation of the two distributions. Equipercentile methods involve matching the distributions at each percentile point.

Pre-test to live test equating

The easiest procedure to carry out is to match the distribution of scores for a pre-test sample (taking the test which will be live in year X+1) with their scores in the live test in year X. Unfortunately this method relies on the assumption that the sample has the same effective distribution of ability in both test situations, and that any differences in raw scores are purely due to differences in the difficulty of the two tests. This is extremely unlikely to be the case because a variety of factors, collectively producing what is known as the 'pre-test effect' will be at work (see, for example, Forster & Bramley, 2002). Chief among these factors are level of motivation (pupils normally will try harder in a live test than in a pre-test), and preparation (the pupils may not have covered, or revised, all the relevant material). Test development agencies attempt to mitigate these two factors by pre-testing as close as possible before the live test, so that all course material will have been covered, and pupils will be well motivated to treat the pre-test as practice for the live test – but nonetheless the pre-test effect cannot be entirely removed and casts doubt on methods based on pre-test to live test equating.

Pre-test to pre-test equating

The main equating method used for KS3 science in the period 1995-2005 has been pre-test to pre-test equating – i.e. matching the distribution of scores in one pre-test sample with the distribution of scores for a pre-test sample in an earlier year. Since the two samples are different, and take different pre-tests, it is possible that the two samples could differ in their distribution of ability and thus it would not be possible to assume that differences in raw score distributions were due to differences in difficulty of the two tests. In fact, by targeting our sample

[9] Test questions covering material in Sc1 (Scientific Enquiry) have only been in the test since 2003.

according to Teacher Assessed (TA) level we aimed to achieve comparable samples from year to year. But a statistical adjustment for differences in ability was made by using an anchor test, a 40-item multiple-choice test[10], which was taken by each pre-test sample.

The equating methods used were those appropriate to an 'anchor test, non-equivalent groups' design (see Petersen et al., 1989). Linear and equipercentile methods were used – my preference was for linear methods where the role of the anchor test was simply to adjust for differences in the mean and SD of the two samples, rather than equipercentile methods where one pre-test was equated to the anchor test, and then the resulting equivalent scores on the anchor test were equated to the second pre-test. The latter approach increases the equating error (see final section) and I also feel that equipercentile equating is likely to 'overfit' the data – in other words to allow for non-linearities which arise mainly because of random fluctuations.

Problems with the anchor test

The pre-test to pre-test methods seemed (in the main) to provide satisfactory cut-scores, and were relatively straightforward to communicate to the client (SCAA / QCA). The only source of contention was the anchor test which, it was suggested, did not measure the same trait as the main test and was also less reliable. These were valid criticisms, especially in 2003 when Sc1 questions came onto the test for the first time (the anchor test only contained questions testing Sc2, Sc3 and Sc4). However, my own view was that these criticisms did not seriously undermine the equating. The role of the anchor test was only to adjust for global differences in the ability of the two samples. The question was merely whether using the anchor test in the statistical method was preferable to not using it. The evidence seemed to show that it was preferable to use it.

Returning to the (dangerous!) analogy with length, the situation was comparable to trying to equate two measuring sticks (with different unknown units) which had measured the heights of two groups of pupils. If we also knew the shoe sizes (in the same units) of these groups of pupils, we could use this information in our equating procedure, even though shoe size is not the same trait as height, and may well not be measured equally reliably. Using the information about shoe size would only be worse than not using it if the relationship between height and shoe size was different in the two groups (e.g. if both groups were on average the same height but one group had on average bigger feet).

In the KS3 science context, it is perhaps arguable that pupils could (over time) get gradually worse at the anchor test whilst not getting any worse (or even improving) at KS3 science generally. But to raise the possibility is not the same as providing evidence that it is true. In any case, even if it were true, it would indicate the direction in which the cut-scores would be likely to err. If the ability of a sample is underestimated by its anchor test performance, the pre-test will appear easier than it really is, and the estimated cut-scores will err on the high side. But in practice it was often the case that doubts about the anchor test were used to cast doubt on the estimated cut-scores whether they appeared too low or too high.

Because of this increasing lack of confidence in the multiple-choice anchor test, in 2006 the cut-scores for KS3 science have been estimated using the same new equating design which was implemented for KS3 English for the first time in 2004, described in section 3.

Vertical equating

Because the KS3 science test has two tiers (targeting pupils at level 3-6 and 5-7 respectively) there also needs to be some attempt at 'vertical equating' within each year, i.e. ensuring that the level 5 and level 6 cut-scores on the two tiers represent the same standard of attainment. This has been done using the common items on the two tiers (90 marks' worth - a substantial portion of each test). 2-stage equipercentile equating of the whole test to the sub-test formed by the

[10] The anchor test used for equating the tests from 1995-2000 contained 16 M-C items. In 2001 there was a separate anchor for Tier 3-6 (32 M-C items) and Tier 5-7 (24 M-C items), with 16 items common to both. In 2002 the two anchor tests were revised and merged to a single test containing 40 M-C items.

common items (as described above with the anchor test) and latent trait (Rasch) methods have both been used and have tended to give similar results. However, when it comes to setting the cut-scores on the live test, historically far more emphasis has been given to ensuring comparability within a tier across years than comparability between the tiers within a year. There is also some evidence to suggest that the relationship between the common items and the whole test is different in a live situation than in a pre-test situation (UCLES, 2004), implying that any vertical equating results obtained from pre-test data should be treated with caution.

3. Equating methods currently used

As mentioned above, a new equating design for the KS3 English test was proposed when the assessment model changed in 2003. The same design is now also used in KS3 Science (from 2006 onwards), so there is now some unity in approach for the two subjects. The new design was described in detail in UCLES (2002). The key feature is the use of a reserve KS3 test as a parallel anchor test. Within each pre-test cohort, one-third of the pupils take version A of the pre-test, one-third take version B, and one-third take the parallel anchor test. The cut-scores on the anchor test are known, so cut-scores on the version A and version B tests can be obtained by simple linear equating or equipercentile equating to the anchor test. This is in effect a 'random groups' design, using the Petersen et al. (1989) terminology, because the three versions of the test are assigned at random within each school that takes part in the pre-test.[11]

There are several advantages to this method:

- Because each new test is equated to the same anchor test there is less possibility for 'drift' in standards which might happen if each year's test is equated to the previous year's test. In my experience of pre-test to pre-test equating there is often some pressure to treat the most recent set of live cut-scores as the most 'correct' set and to use those as the reference year. This has the potential to accumulate equating error. Using the same anchor test means that (in theory) the cut-scores are as likely to err on the high side as on the low side each year.
- The new anchor test is a completely parallel test, written to the same specification and pre-tested with the same rigour as a live test. Obviously after a long time period (5 years?) it may start to look dated but as long as the test specification does not change there can be no doubt about its relevance.
- The fact that the equating is done within a single pre-test year group means that most aspects of the pre-test effect are controlled to a greater extent than previously: there are no school effects (differences arising from different teachers and pupils in different schools), no motivation effects – the level of motivation for pre-test and anchor test group should be the same (the pupils are unaware which is the anchor test and which is the 'real' pre-test), and no preparation effects – any lack of preparation which affects scores on the pre-test should affect scores on the anchor test equally, since it is a parallel test.

Apart from the possibility that the anchor paper may become 'out of date', the only disadvantage to this method is the cost of pre-testing the extra paper – the sample size needs to be increased by 50%, with a consequent increase in printing, invigilation and marking expense.

In KS3 English, there are two extra complicating factors:

Separate cut-scores for Reading and Writing

Cut-scores have to be derived for both Reading and Writing (the sum of these cut-scores provides the overall cut-score for KS3 English). The fact that equating is now done twice (once for Reading and once for Writing) and the results added together means that any equating error is doubled. The splitting of the test into two components (Reading and Writing) halves the length

[11] In KS3 English the papers are pre-allocated to ensure an equal distribution of pupils at each TA level within each school – this is even better than random allocation in ensuring equivalent groups. In KS3 science the papers are 'spiralled' – i.e. version A, version B and the anchor are delivered to every third pupil. This results in randomly equivalent groups provided there is no ability-related pattern in the seating plan.

of the mark scale from 100 to 50, which reduces the reliability of the scores (i.e. increases the proportion of variance due to measurement error). It also means that the effects of rounding the equated cut-scores to the nearest whole number are bigger, because a larger proportion of pupils will obtain each mark on a shorter mark scale than a longer one. In the middle of the score ranges for Reading and Writing there might be as many as 4% of the pupils on each mark. Thus the choice between two cut-scores one mark apart could have an effect of 4% on the proportion of pupils reaching the level. Given that this choice may depend on whether an equated cut-score was rounded up or down, and given the earlier comments about measurement error and equating error, it is clear (to me) that statistical equating for KS3 English is not likely to give the precision required for the political purposes of monitoring changes in the proportion of pupils at each level. The problem is less acute at KS3 Science because the mark scales are longer (180 marks for T3-6 and 150 marks for T5-7).

Shakespeare

Part of the Reading component of the KS3 English test comes from the 18-mark Shakespeare task. There is no really defensible way to allow for variations in difficulty of this paper from year to year. First of all, the Shakespeare task does not form part of the anchor test, because the set scenes change each year (and choice of plays changes on a rolling cycle). It is also difficult to recruit schools willing to pre-test the task – because it will be on a different scene from the one which their pupils will be preparing for the current year's test. Even if schools do prepare their pupils for an extra scene, it is hard to generalise from their performance in these unusual conditions to what the general cohort will do in a live test.

This means that the statistical equating for Reading is based on cut-scores derived for the 32-mark Reading comprehension paper (which can be equated to the anchor test), plus some allowance for the contribution of the Shakespeare Reading task. The simplest solution, and probably the most defensible, has been to add a constant number of marks (a different value at each level), based either on the average mark on the Shakespeare task obtained in the past by pupils on each level boundary in the total cohort, or on the average mark at each level which would have been needed in the past to get from equated cut-score on the 32-mark test to final cut-score on the 50-mark Reading component. Both these methods effectively assume that the Shakespeare task is equally difficult each year. The same assumption needs to be made about the three tasks on the different plays within a year, although in this case information from pre-testing can be more validly used to highlight tasks of different difficulty which can then be modified.

In my view the presence of the Shakespeare task further muddies what was already fairly murky water. To improve statistical equating in KS3 English I would recommend removing the Shakespeare element from the test and increasing the Reading comprehension test from 32 to 50 marks, as at KS2.

4. Issues in equating and standard maintaining

Changes to the assessment model

It is especially difficult to equate when the assessment model (or test specification) changes because the presence of new or different content and/or item types means that it is less safe to assume that the two tests are measuring the same trait. With a pre-test to pre-test equating situation it is also less safe to assume that the anchor test is equally relevant to the old and the new test. Finally, arguments about the extent to which schools have been able to prepare their pupils for the new test often affect the setting of the cut-scores. If there is an emphasis on 'being fair to this year's pupils' in the year of the new test (by allowing for the fact that teachers will have been less prepared for it by setting lower cut-scores than might otherwise have been set), followed by 'rewarding the improved performance' of the subsequent year's pupils (whose teachers have become accustomed to the new system) then this provides a mechanism for pass rates to rise without any genuine improvement in the level of the trait in the cohort, as was explained by Pollitt (1998).

Perhaps the simplest way to deal with this problem is by specifying in the level-setting procedures some kind of rule to the effect that when the test or specification changes significantly, cut-scores are set such that there is a reduction of (say) 2% in the cumulative percentage of pupils at each level.

Equating error

The statistical methods which are used for equating are subject to random error due to sampling fluctuations. It is possible to estimate an 'equating error' for both equipercentile and linear equating methods. This error can be interpreted as the standard deviation of equated scores on test X corresponding to a particular score on test Y if the equating procedure was carried out many times with different random samples from the same population (Petersen et al., 1989). The equating error will often be different at different points on the mark scale (usually higher at greater distances from the mean).

We have tended not to report equating error when presenting our analyses. This is partly because the calculations require yet further assumptions which can be hard to justify, but mainly because there is a danger that the resulting 'confidence interval' for the cut-score would be interpreted as the range in which the 'true' cut-score lies, after allowing for *all* the various assumptions and caveats involved (see section 1) rather than simply random sampling error. However, it is arguably better practice to report an estimate of equating error nonetheless, and it would be desirable to involve it in a rational way in the level setting process. A new suggestion for how to do this is outlined below.

A Bayesian method for combining different sources of evidence when setting cut-scores

In my opinion the biggest problem with maintaining standards in National Tests is the lack of a clearly specified rationale for combining the various sources of evidence when setting the final cut-scores. The political importance which is attached to the percentage of pupils at each level means that the 'impact data' (the distribution of scores on a nationally representative large sample of 20,000+ pupils) which is available at the final level setting meeting can dominate the decision-making process. The underlying assumption seems to be that the distribution of levels should not change greatly from one year to the next, and, if it does change, an improvement is more credible than a decline. This assumption is arguably very reasonable, but it clearly cannot be used as a basis for setting cut-scores in a system designed to *measure* changes over time.

The approach I am about to propose will not address the problem of monitoring changes over time, which many experts agree is best approached by light sampling methods (e.g. Johnson, 1988). It merely attempts to formalise the role which expectations about changes in the distribution of levels in the population could have in the cut-score setting procedure.

I have often wondered whether some Bayesian-inspired method for combining a distribution of 'prior beliefs' about the likelihood of various percentage rises and falls with a probability distribution for the location of the cut-score (derived from statistical equating) might provide a rational basis for setting the live cut-scores. Such a system might work as follows:

Step 1: QCA and the DfES supply *in advance* their 'official' probabilities for a range of possible changes in the cumulative percentage of pupils at a level (I will assume level 5 in the example below – in theory they could specify a different set of prior probabilities at each level).

Table 1. Hypothetical prior probability distribution for changes in the cumulative percentage of pupils at Level 5.

Change in %	≤-5%	-4%	-3%	-2%	-1%	0%	1%	2%	3%	4%	5%	≥6%
Probability	0	0.01	0.03	0.06	0.15	0.25	0.25	0.15	0.06	0.03	0.01	0

Table 1 quantifies the assumption (prior expectation) that there is no possibility that the cumulative percentage of pupils at level 5 could fall by more than 4 percentage points, and no possibility of it rising by more than 5 percentage points. Within this range, the probabilities are symmetrically distributed with the most probable changes being 0 and +1.

Step 2: The test development agency supplies a range of cut-scores based on the statistical equating with a probability attached to each – this would provide an intuitive way to allow for equating error.

Table 2: Hypothetical range of values for the KS3 Science level 5 cut-score with a probability attached to each.

L5 cut-score	≤101	102	103	104	105	106	107	108	109	110	111	≥112
Probability	.001	.004	.018	.054	.119	.192	.225	.192	.119	.054	.018	.005

The values in Table 2 are based on an equated cut-score of 107 with an equating error of 1.75 marks (a plausible value for the middle of the score range in KS3 science). This error is assumed to be normally distributed in order to calculate the probabilities. Thus, purely on the basis of the statistical equating, the cut-score in this example is thought most likely to be 107, and with most of the probability spread across the range 105-109.

Step 3: Once the live test data from the nationally representative sample of 20 000+ pupils has been collected, the cumulative percentage of pupils at each mark on the test is calculated. The range of scores on the test corresponding to the ‘acceptable’ rise or fall in cumulative percentage is determined, using the information in Table 1 combined with knowledge of the previous year’s cumulative percentage of pupils at the Level 5 cut-score. For the sake of simplicity, Table 3 assumes that exactly the same probabilities as in Table 1 apply to the range of consecutive marks, though in practice this could only occur if exactly 1% of the pupils were on each successive mark in the range of interest.

Table 3: Combining expectations about pass rates with equating results to derive the cut-score.

Mark	Prior probability $p(H)$[12]	Likelihood of equating result $p(D\|H)$	Posterior probability $p(H\|D)$
97	0	≈0	
98	0.01	≈0	
99	0.03	≈0	
100	0.06	≈0	
101	0.15	0.001	0.004
102	0.25	0.004	0.037
103	0.25	0.018	0.153
104	0.15	0.054	**0.280**
105	0.06	0.119	0.248
106	0.03	0.192	0.200
107	0.01	0.225	0.078
108	0	0.192	
109	0	0.119	
110	0	0.054	
111	0	0.018	
112	0	0.004	

The information in Table 3 can be interpreted as follows:

[12] I have followed the convention of using H for hypothesis and D for data. $p(H=x \mid D)$ is thus the posterior probability that the cut-score = x, given the observed equating data D.

The first column contains the marks on the test in the range considered for the Level 5 cut-score. The second column contains the 'prior probabilities' for each of these cut-scores, determined in advance (and, I am suggesting, enshrined in a written procedure!), applied to these particular marks using information about the cumulative distribution of the National Sample. The second column shows that a cut-score of 103 would give the same cumulative percentage at level 5 as the previous year, and that a cut-score of 102 would give a rise of 1% in the cumulative percentage. (Recall that these were the two outcomes with the highest 'prior probability' in Table 1).
The third column repeats vertically the information in Table 2, showing the probability of each cut-score based on the statistical equating. A cut-score of 107 is the most probable, based solely on this evidence.
The fourth column contains the 'posterior probabilities' of each cut-score, based on combining the probabilities[13] in columns 2 and 3 according to Bayes' formula:

$$p(H = x \mid D) = \frac{p(D \mid H = x) \times p(H = x)}{\sum_{x=97}^{x=112} p(D \mid H = x) \times p(H = x)}$$

This shows how prior expectations can be combined with evidence from equating, allowing for equating error. With this (purely hypothetical) data, the resulting cut-score could either be taken as the one with the highest posterior probability, (i.e. 104), or the mean of the posterior distribution (in this case 104.6, which would be rounded to 105). Thus it can be seen that the equating evidence would 'pull' the cut-score up, but that the prior expectations prevent it from moving too far away from a mark which would give an 'acceptable' outcome – not an unreasonable model for what actually happens in practice! This model could possibly be extended to include evidence from script scrutiny and/or judgmental exercises.

Of course, a key part of this approach would be in producing the 'prior distribution' reflecting the beliefs / assumptions / expectations of the relevant people. Deciding who are the relevant people would also be important! This prior distribution could either be specified directly, or elicited by some kind of judgmental exercise. A 'sensitivity analysis' could be carried out to discover how robust the outcome is to variations in the shape of the prior distribution.

The above is at this stage by no means a fully worked-out system, but just an idea which I put forward in the hope it will provoke some interesting debate. I feel that the main two advantages it might offer are:

- a natural way to allow for 'equating error'
- a formal way to involve expectations about likely changes in the proportion of pupils at each level, which should allow conflicts in the various sources of evidence about the cut-scores to be handled more smoothly in the final level setting process.

[13] The zeros in columns 2 and 3 of Table 3 could be replaced with very small numbers so that every mark has a non-zero posterior probability, but this would have no effect on the overall outcome in this example.

References

Bramley, T. (2005a). A rank-ordering method for equating tests by expert judgment. *Journal of Applied Measurement,* 6 (2) 202-223.

Bramley, T. (2005b). Accessibility, easiness and standards. *Educational Research, 47 (2), 251-261.*

Borsboom, D., Mellenbergh, G.J., & Van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, 110, 203-219.

Forster, M. & Bramley, T. (2002). An investigation of the pre-test effect. Report to QCA.

Johnson, S. (1988). *National Assessment: the APU science approach*. London: APU.

Lord, F. M. (1952). *A theory of test scores.* New York: Psychometric Society.

Massey, A.J. (1995). Criterion-related test development and national test standards. *Assessment in Education,* 2, 2, 187-203

Massey, A., Green, S., Dexter, T. and Hamnett, L. (2003). *Comparability of national tests over time: KS1, KS2 and KS3 standards between 1996 and 2001.* QCA. London.

Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology, 10(5)*, 639-667.

Newton, P.E. (2005). Examination standards and the limits of linking. *Assessment in Education, 12 (2),* 105-123.

Pelton, T.W. & Bunderson, C.V. (2003). The recovery of the density scale using a stochastic quasi-realisation of additive conjoint measurement. *Journal of Applied Measurement 4(3)*, 269-281.

Petersen, N.S., Kolen M.J., and Hoover H.D. (1989) Scaling, Norming and Equating. In R.L.Linn (Ed.), *Educational Measurement* (3rd ed.), Phoenix, Arizona: The Oryx Press.

Pollitt, A. (1994). Standards in KS3 English. Report to SCAA.

Pollitt, A. (1998). Maintaining standards in changing times. Paper presented at the 24th annual conference of the International Association for Educational Assessment (IAEA), Barbados.

Stenner, A.J. (1996). Measuring reading comprehension with the Lexile framework. Paper presented at the California Comparability Symposium, Burlingham, CA.

UCLES (2002). Key Stage 3 English: equating strategies in transition 2002-2004. Report to QCA.

UCLES (2004). NAA Key Stage 3 Science: between tier comparability. Report to QCA.

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement*, 14 (2), 97-116.